软件所四项成果被自然语言处理国际会议ACL2019接收

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/5222.html

软件所四项成果被自然语言处理国际会议ACL2019接收。ACL(Annual Meeting of the Association for Computational Linguistics)是自然语言处理领域的顶级国际会议，被中国计算机学会推荐国际学术会议列表认定为A类会议。ACL2019将于2019年7月28日至8月2日在意大利佛罗伦萨举行。

在国家自然科学基金重点项目“汉语认知加工机制与计算模型”和国家重点研发计划“基于大数据的面向开放域的智能问答技术”项目支持下，中国科学院软件研究所中文信息处理实验室3篇长文Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks、Distilling Discrimination and Generalization Knowledge for Event Detection via ?-Representation Learning、Progressively Self-Supervised Attention Learning for Aspect-Level Sentiment Analysis和1篇短文Cost-sensitive Regularization for Label Confusion-aware Event Detection 被ACL2019接收。

(1) Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks

论文作者：林鸿宇(中科院软件所)，陆垚杰(中科院软件所)，韩先培(中科院软件所)，孙乐(中科院软件所)

命名实体识别是自然语言处理中一个根本性的任务。然而，现有的命名实体识别模型通常假定一个字符仅属于一个实体提及，这就使得这些模型无法被用于带有嵌套命名实体提及的情况。但是，嵌套命名实体提及在自然语言中分布非常广泛，这就使得忽视这类嵌套实体会对后续自然语言处理任务产生巨大影响。

针对这一问题，该团队提出了一种全新的神经网络结构：锚点-区域网络。该网络充分地利用了自然语言词组以头词为中心的结构特性，提出了通过检测不同头词来检测不同嵌套实体的方案。同时，为了能够在没有实体头词标注数据的情况下训练上述网络结构，团队还提出了一种新的包损失函数。该损失函数能够自动挖掘无头词标注数据中的头词信息，从而对锚点-区域网络进行端到端训练。

实验结果表明该团队提出的模型在ACE2005、GENIA以及KBP2017等多个不同领域的命名实体识别标准数据集上都取得了当前最好的性能。

(2) Distilling Discrimination and Generalization Knowledge for Event Detection via Representation Learning

论文作者：陆垚杰(中科院软件所)，林鸿宇(中科院软件所)，韩先培(中科院软件所)，孙乐(中科院软件所)

事件检测是信息抽取的重要任务，近年来，在知识图谱构建、信息检索和文本理解中扮演着重要角色。事件检测系统不仅依赖判别性知识来区分存在歧义的事件触发词，还依赖泛化性知识来检测未见的、稀疏的事件触发词。现有的神经网络方法通常聚焦于获取一个以触发词为中文的特征表示用于事件检测，这样的方法可以有效地蒸馏出判别性的知识，但是难以学习到泛化性的知识，致使模型难以检测未见的、稀疏的事件触发词。

为解决这一问题，论文提出了一种表示学习框架，通过有效分离、增量学习，最后自适应合成不同的事件特征表示，能够有效地蒸馏判别性和泛化性知识。

实验结果证明了该文的方法在未见的、稀疏的事件触发词上超过了之前的方法，同时在ACE2005和KBP2017两个数据集取得了当前最佳性能。

(3) Progressively Self-Supervised Attention Learning for Aspect-Level Sentiment Analysis

论文作者：唐家龙(中科院软件所)，陆紫耀(厦门大学)，苏劲松(厦门大学)，葛毓斌(UIUC)，宋霖峰(罗切斯特大学)，孙乐(中科院软件所), 罗杰波(罗切斯特大学)

在方面层次的情感分类任务中，使用注意力机制来捕获上下文文本中与给定方面最为相关的信息是近年来研究者们的普遍做法。然而，注意力机制容易过多地关注数据中少部分有强烈情感极性的高频词汇，而忽略那些频率较低的词。

该文提出了一种渐进的自监督注意力的学习算法，能够自动、渐进地挖掘文本中重要的监督信息，从而在模型训练过程中约束注意力机制的学习。该团队迭代地在训练实例上擦除对情感极性“积极”/“消极”的词汇。这些词在下一轮学习过程中将会被一个特殊标记替代，并记录下来。最终，团队针对不同情况，设计出不同的监督信号，在最终模型训练目标函数中作为正则化项约束注意力机制的学习。

在SemEval 14 REST，LAPTOP以及口语化数据集TWITTER上的实验结果表明，团队提出的渐进注意力机制能在多个前沿模型的基础之上取得显著性提升。

(4) Cost-sensitive Regularization for Label Confusion-aware Event Detection

论文作者：林鸿宇(中科院软件所)，陆垚杰(中科院软件所)，韩先培(中科院软件所)，孙乐(中科院软件所)

事件检测是信息抽取中的一个重要任务。近年来，神经网络在事件检测上取得了重大的进展。然而，该研究发现，神经网络模型在事件检测上的错误通常出现在某些特定的类别对之间。针对上述问题，研究人员提出了一种代价敏感的正则化约束优化目标。该约束目标使得神经网络在训练的过程中能够更加关注某些特定的易混淆类别对。除此之外，他们还提出了两种实例级别以及语料库级别的用于估计类别间混淆度的方法。在ACE2005以及KBP2017数据集上实验结果表明，他们提出的代价敏感的正则化约束能够显著提升多种不同架构的神经网络事件检测模型的性能。

论文全文和源代码将在中文信息处理实验室网(www.icip.org.cn)开放。